Nuclear Physics B203 (1982) 205–220
© North-Holland Publishing Company

't HOOFT'S CONSISTENCY CONDITION AS A CONSEQUENCE OF ANALYTICITY AND UNITARITY

Sidney COLEMAN

Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

Bernard GROSSMAN

Department of Physics, Rockefeller University, New York, New York 10021, USA

Received 7 April 1982

We derive 't Hooft's consistency condition on the bound-state spectrum of a confining field theory from the principles of analytic S-matrix theory.

1. Introduction

In his 1979 Cargèse lectures [1], 't Hooft advanced a consistency condition for the spectrum of a confining field theory. This paper gives an argument for this condition based on analyticity and unitarity, that is to say, upon the principles of analytic S-matrix theory developed in the 1960's [2].

We do not claim to say here either the first or last word on this subject. In his original paper, 't Hooft gave an argument for the condition which many physicists (including us) find convincing; shortly afterwards, Frishman et al. [3] gave a derivation along dispersive lines, which shares common ideas with our work. Also, there are several places in our argument where we have had to make assumptions which we suspect are unnecessary but which we have not been able to eliminate. Nevertheless, we believe we have made a step forward, that our argument simplifies reasoning that was formerly complex and clarifies points that were formerly obscure.

We begin by stating the consistency condition; this requires some preliminary definitions. To keep things as simple as possible, we will restrict ourselves here to chromodynamics with color group SU(N) and with n flavors of massless quarks. The generalization to more complex theories is straightforward. We represent the quarks by Dirac fields ψ^{ar}, where $a = 1 \ldots n$ is the flavor index and $r = 1 \ldots N$ is the color index. Let J^μ be some colorless flavor current,

$$J^\mu = \bar{\psi}^{ar}\gamma^\mu[A_+^{ab}(1+\gamma_5)+A_-^{ab}(1-\gamma_5)]\psi^{br}, \tag{1.1}$$

where the A's are some hermitian matrices, and the sum on repeated indices is

[*] The viewpoint that the anomaly gives rise to a singularity at $Q^2 = 0$ was first discussed in the context of perturbation theory by Dolgov and Zakharov [3].

implied. This current will be conserved,

$$\partial_\mu J^\mu = 0 , \tag{1.2}$$

if it is free of chromodynamic anomalies,

$$\mathrm{Tr}\,(A_+ - A_-) = 0 . \tag{1.3}$$

We will restrict ourselves to currents for which this is so.

Of course, we do not expect the spectrum of the theory to consist of quarks and gauge mesons, but rather of colorless particles, hadrons. Let us focus attention on the massless spin-$\frac{1}{2}$ hadrons; we denote these states (relativistically normalized) by $|u, p, \alpha\rangle$, where u is a solution of the massless Dirac equation, p is a null four-momentum, and α labels the hadron. We define matrices $B_\pm$ by

$$\langle u', p, \alpha | J^\mu(0) | u, p, \beta\rangle = \bar{u}'\gamma^\mu[B_+^{\alpha\beta}(1+\gamma_5) + B_-^{\alpha\beta}(1-\gamma_5)]u . \tag{1.4}$$

We can now state the consistency condition. If the symmetry associated with J^μ does not suffer spontaneous breakdown, then

$$\mathrm{Tr}\,(B_+^3 - B_-^3) = N\,\mathrm{Tr}\,(A_+^3 - A_-^3) . \tag{1.5}$$

(There are also non-trivial results if the symmetry does suffer spontaneous breakdown, but they have to do with the properties of Goldstone bosons, not of spin-$\frac{1}{2}$ hadrons.)

Eq. (1.5) is clearly connected with the anomalous Ward identity first discovered by Adler, Bell, and Jackiw [4]. Let us define

$$\Gamma_{\mu\nu\lambda}(q_1, q_2, q_3)(2\pi)^4\delta^{(4)}(q_1+q_2+q_3) = \int \prod_i \mathrm{d}^4x_i\, \mathrm{e}^{iq_i\cdot x_i} T < 0|J_\mu(x_1)J_\nu(x_2)J_\lambda(x_3)|0\rangle . \tag{1.6}$$

Then

$$q_3^\lambda \Gamma_{\mu\nu\lambda} = \frac{N}{\pi^2}\mathrm{Tr}\,(A_+^3 - A_-^3)\varepsilon_{\mu\nu\alpha\beta}q_1^\alpha q_2^\beta . \tag{1.7}$$

We will refer to this as "the anomaly equation" and to its right-hand side as "the anomaly". The anomaly equation is known to be true to all orders of perturbation theory; we will assume here that it is true exactly. In an appendix we give a hand-waving argument that non-perturbative effects (e.g., instantons) do not change the anomaly.

(The usual form of the anomaly equation involves three distinct currents, and the reader may think we have thrown away valuable information by restricting ourselves to three identical currents. This is not so. The part of the anomaly that is not totally symmetric under the interchange of the three currents can always be made to vanish by an appropriate choice of the subtraction constants in the definition

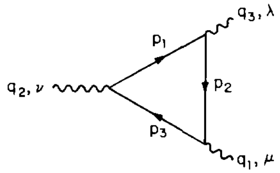

Fig. 1. The triangle graph.

of the T-product; only the totally symmetric part is convention-independent and only this contains genuine dynamical information*.)

Despite the fact that Γ has non-trivial contributions in all orders of perturbation theory, eq. (1.7) is independent of the coupling; the anomaly is given totally by the zeroth-order contribution, the triangle graph of fig. 1. This leads to a striking formulation of the consistency condition: If one treats the massless spin-$\frac{1}{2}$ hadrons as if they were fundamental particles, and if one ignores all other particles, one nevertheless correctly computes the anomaly. (This is essentially 't Hooft's original formulation.)

This way of putting things leads to some obvious questions:

(1) We normally think of the anomaly equation as telling us something about the fundamental fields in the theory, about its short-distance structure. Thus, naively, we would expect it to lead to a sum rule involving contributions from particles of every mass. Why are only massless particles important?

(2) Hadrons are not elementary particles, and their interactions need not be governed by the Feynman rules of a renormalizable field theory. How does the triangle graph enter the picture, and why is it the only important graph?

(3) What is special about spin $\frac{1}{2}$? Why is there no contribution from massless particles of other spins**?

* *Proof:* Let us label the currents by indices a, b, c. In general the anomaly can be of the form

$$q_3^\lambda \Gamma^{abc}_{\mu\nu\lambda} = a^{abc} \varepsilon_{\mu\nu\alpha\beta} q_1^\alpha q_2^\beta \, .$$

The only conditions on the numbers a^{abc} given by Bose symmetry is $a^{abc} = a^{bac}$. We are always free to redefine Γ by adding a term that vanishes except when the three currents are at the same space-time point. In momentum space, this corresponds to adding a polynomial in momenta. Let us make the redefinition

$$\Gamma^{abc}_{\mu\nu\lambda} \to \Gamma^{abc}_{\mu\nu\lambda} - \tfrac{1}{3}\varepsilon_{\mu\nu\lambda\sigma}[(a^{abc} - a^{cab})q_1^\sigma + (a^{bca} - a^{abc})q_2^\sigma + (a^{cab} - a^{bca})q_3^\sigma] \, .$$

Note that the added term preserves the full Bose symmetry of Γ under permutations of the three currents. Under this redefinition,

$$a^{abc} \to \tfrac{1}{3}(a^{abc} + a^{bca} + a^{cab}) \, .$$

This is fully symmetric under permutation of the indices.

** In fact, as Weinberg and Witten [5] have shown, there are severe restrictions on possible massless-particle helicities in chromodynamics. However, we will not need to use their results here. If we were to use them, we would simplify our work only slightly; for example, we would still have to worry about possible helicity one-helicity zero transition vertices in sect. 4.

In the remainder of this paper, we shall develop our argument and answer these three questions in the order in which we have posed them. In sect. 2, we shall show that the anomaly equation implies a singularity in an invariant amplitude at zero momentum transfer; such a singularity is given exclusively by the contribution of massless particles. In sect. 3, we shall review the rules for locating singularities in the physical region and the rules for computing the discontinuities at such singularities, the Cutkosky rules. The triangle graph will emerge not as a Feynman graph but as a generalized unitarity graph. Finally, in sect. 4, we shall compute the residue of the singularity including all spin factors. Our expression will involve kinematic zeros computable in terms of the helicities of the particles going around the triangle. These kinematic zeros will turn out to kill the singularity unless the helicities are all equal to $\pm\frac{1}{2}$. In sect. 5, we discuss ways in which our arguments might be improved.

2. Analyticity and the importance of massless hadrons

Γ is completely symmetric under simultaneous permutations of (q_1, q_2, q_3) and (μ, ν, λ). The restrictions this places on its expansion in terms of invariant amplitudes is strongest along the line

$$q_1^2 = q_2^2 = q_3^2 \equiv -Q^2 . \tag{2.1}$$

We will restrict ourselves to this line here. Among its other advantages, this policy saves us from the horrors of the theory of functions of several complex variables.

The covariants occurring in the expansion of Γ can be either tensorial or pseudotensorial (i.e., involving the ε-symbol). Only the latter can contribute to the anomaly, so we will restrict ourselves to studying these. (In fact, the tensorial invariants vanish, but this information is not necessary for our argument.) At least one of the four indices of the ε-symbol must be contracted with a momentum vector. Because the sum of the three-momenta vanishes, there are only two independent vectors in the problem, and at most two of the indices can be contracted with momentum vectors.

Thus we have only two cases to consider.

(1) One index is contracted. For example,

$$\Gamma_{\mu\nu\lambda} = F(Q^2)\varepsilon_{\mu\nu\lambda\alpha}q_1^\alpha + \cdots , \tag{2.2}$$

where F is some unknown function. However, if we symmetrize under cyclic permutations, this becomes

$$F(Q^2)\varepsilon_{\mu\nu\lambda\alpha}(q_1^\alpha + q_2^\alpha + q_3^\alpha) = 0 . \tag{2.3}$$

(2) Two indices are contracted. Let us focus on the terms in which μ and ν are the uncontracted indices; the others are determined by these through the permutation symmetry. Because there are only two independent vectors in the problem,

there is no choice for which vectors to contract; however, there are two choices for which vector carries the remaining index, λ. We can choose the two independent possibilities to be

$$\Gamma_{\mu\nu\lambda} = F(Q^2)\varepsilon_{\mu\nu\alpha\beta}q_1^\alpha q_2^\beta(q_1 - q_2)_\lambda + \cdots, \tag{2.4}$$

$$\Gamma_{\mu\nu\lambda} = F(Q^2)\varepsilon_{\mu\nu\alpha\beta}q_1^\alpha q_2^\beta q_{3\lambda} + \cdots. \tag{2.5}$$

The first of these is odd under simultaneous interchange of (q_1, q_2) and (μ, ν), and thus is not acceptable. Fortunately, the second of these can be successfully symmetrized; it is the only allowed pseudotensorial covariant.

Thus,

$$\begin{aligned}\Gamma_{\mu\nu\lambda} = F(Q^2)[&\varepsilon_{\mu\nu\alpha\beta}q_1^\alpha q_2^\beta q_{3\lambda} \\ &+ \varepsilon_{\nu\lambda\alpha\beta}q_2^\alpha q_3^\beta q_{1\mu} + \varepsilon_{\lambda\mu\alpha\beta}q_3^\alpha q_1^\beta q_{2\nu}] \\ &+ \text{possible tensors}.\end{aligned} \tag{2.6}$$

The anomaly equation, eq. (1.7), becomes

$$-Q^2F(Q^2) = \frac{N}{\pi^2}\operatorname{Tr}(A_+^3 - A_-^3).$$

That is to say, F is a simple pole in Q^2, with known residue.

We now see how to reconcile the two faces of the anomaly. The anomaly equation gives us the form of F for *all* Q^2. Thus, at one and the same time, it gives us the leading asymptotic form of F for large Q^2, short-distance behavior, and the leading singularity for small Q^2, large-distance behavior. To make this last observation more precise we will have to study the rules that give us the singularities of Feynman amplitudes.

3. Singularities and the importance of traingles

For the next stage of our investigation, we need two tools of S-matrix theory, the physical-region Landau rules and the Cutkosky rules. These tell us when an amplitude has a singularity in the physical region, the region of real external momenta, and how to compute the discontinuity at the singularity. These rules were first derived in perturbation theory; however, there are arguments (though not rigorous ones) that they should be true independent of perturbation theory. We will simply state them here and assume their validity; the reader interested in their derivation is referred to the literature [3].

The physical-region Landau rules state that a singularity occurs only for those external momenta for which one can draw a graph of a process contributing to the amplitude such that (1) all internal particles are stable particles on the mass shell, (2) all vertices conserve energy and momentum, and (3) the graph can be interpreted

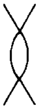

Fig. 2. A generalized unitarity graph which gives a singularity of normal-threshold type in two-into-two scattering. The two internal lines are bowed apart from each other for clarity in the drawing; as argued in the text, in fact they coincide.

as a drawing of a process going on in space–time, with all internal particles obeying the rules of classical kinematics for free motion and moving forward in time. These graphs look something like Feynman graphs, but they are not Feynman graphs; the first and third conditions are very unlike any for Feynman graphs, and even in a renormalizable theory, the vertices may have as many particles as one wants coming together at them. We will call these graphs generalized unitarity graphs. (In some formulations of S-matrix theory, these rules are derived from a generalization of the unitarity equation.)

As an example, consider the graph for a two-particle scattering process shown in fig. 2. (The various particles need not be identical.) We argue that the two internal particles must be at rest in the c.m. frame, for if they had non-zero velocities, they would move apart from each other from their moment of production, and thus, by classical kinematics, could never join each other at the second vertex. Thus the singularity is at the two-particle production threshold.

The Cutkosky rules state that to determine the discontinuity associated with such a graph, we evaluate it as if it were a Feynman graph, except that every internal line carries, instead of a Feynman propagator, a factor of $2\pi i\theta(p^0)\delta(p^2-m^2)$, and that each vertex is to be replaced by the exact amplitude for the relevant process.

It is important to realize that these rules give us the discontinuities of the full scattering amplitudes (or Green functions) not of the scalar amplitudes into which they are expanded. As we shall see, this can cause difficulties if we are not careful.

It will be convenient for us to use what are sometimes called light-cone basis vectors,

$$\begin{aligned} e_1 &= (0,1,0,0)\,, \\ e_2 &= (0,0,1,0)\,, \\ e_\pm &= \sqrt{\tfrac{1}{2}}(1,0,0,\pm 1)\,. \end{aligned} \tag{3.1}$$

A natural way to satisfy eq. (2.1) would be to construct the three external momenta by taking a triplet of spacelike unit vectors that sum to zero and multiplying them by Q. For example, we could choose

$$\begin{aligned} q_1 &= Qe_1\,, \\ q_2 &= \tfrac{1}{2}Q(-e_1+\sqrt{3}e_2)\,, \\ q_3 &= \tfrac{1}{2}Q(-e_1-\sqrt{3}e_2)\,. \end{aligned} \tag{3.2}$$

This choice would be a disaster, though, because

$$\mathrm{disc}\,\Gamma \propto Q^2\delta(Q^2) = 0\,. \tag{3.3}$$

This problem is easily circumvented. We choose

$$\begin{aligned} q_1 &= Qe_1 + e_+\,, \\ q_2 &= \tfrac{1}{2}Q(-e_1 + \sqrt{3}e_2)\,, \\ q_3 &= \tfrac{1}{2}Q(-e_1 - \sqrt{3}e_2) - e_+\,. \end{aligned} \tag{3.4}$$

Eq. (2.1) is still satisfied, and now

$$\mathrm{disc}\,\Gamma \propto Q\delta(Q^2) \neq 0\,. \tag{3.5}$$

Eq. (3.4) shows that only very special graphs can produce a singularity at $Q = 0$. There is only one vertex at which positive energy is injected into the graph, the q_1 vertex, so this must be the first vertex in space-time. For $Q = 0$, q_1 is a null vector, so the particles produced at this vertex must be all massless and all collinear. As time goes on, these particles may interact with each other at internal vertices of the graph, or interact with an external current (carrying zero momentum) at the q_2 vertex. However, the total momentum of the system remains null; thus, though the particles may change their character, number, and helicity, they must remain massless and collinear. Finally, all of the particles must be absorbed at the q_3 vertex.

The simplest graph fitting this description is the triangle graph of fig. 1. Actually, as a space–time picture, fig. 1 is badly drawn; the broken line labeled by p_3 and p_1 should form a single straight line coincident with the p_2 line. We will study this graph more closely before going on to other possibilities.

We know the triangle graph has a singularity in the right place, $Q^2 = 0$, but is it the right singularity? To get a preliminary idea of what is going on, let us simplify the problem by ignoring all tensorial structure and also ignoring all possible momentum dependence of the verticles in the graph. The Cutkosky rules would then predict a discontinuity proportional to the integral

$$I = \int \mathrm{d}^4p_1 \prod_{r=1}^{3} \theta(p_r^0)\delta(p_r^2)\,. \tag{3.6}$$

This integral can be evaluated with no work. By the space-time interpretation of the triangle graph, q_1 must be timelike or null and q_2 must be spacelike or null; thus the support of the integral must be restricted to $Q^2 = 0$. By dimensional analysis, the integral is a homogeneous function of Q of order minus two. Thus,

$$I \propto \delta(Q^2)\,. \tag{3.7}$$

This is very close to the desired result, eq. (3.5). As we shall see in sect. 4, the vertices are in fact momentum-dependent, and this supplies the missing factor of Q.

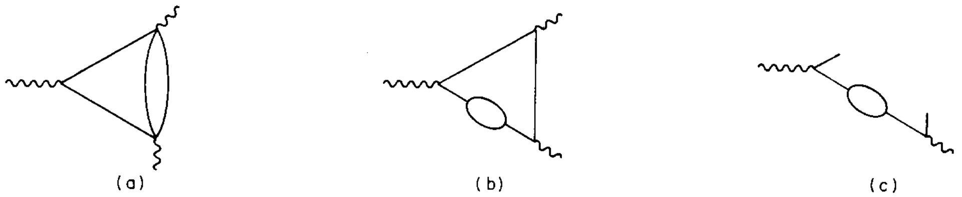

Fig. 3. Three generalized unitarity graphs. In all cases, time runs upwards. As in fig. 2, coincident lines are separated for clarity. (c) is just the bottom half of (b).

This kind of argument can be used to discard whole families of graphs. For example, fig. 3a also obeys the physical-region Landau rules and could produce a singularity at $Q^2=0$. However, dimensional analysis here leads to a discontinuity that is a homogeneous function of order zero, and there is no such function with support restricted to a single point. Of course, we are tacitly assuming here that the vertices themselves are free of singularities. We will continue to make this assumption throughout the remainder of our argument; we will examine it in more detail at the end of the paper.

Unfortunately, this argument does not eliminate all graphs. For example, dimensional analysis applied to fig. 3b leads to a discontinuity of order minus four, that is to say, to a double pole in the amplitude. Fortunately, we can show that such a singularity is not possible. If there is a double pole in fig. 3b, there is correspondingly a double pole in the amplitude for two currents make two mesons, fig. 3c. But such a singularity in momentum space would lead to an amplitude in position space that grew logarithmically with separation for large separation of the currents. This is in contradiction to the general principles of field theory. Thus the potential double pole in fig. 3c, and correspondingly in fig. 3b, must be damped by momentum factors at the vertices.

We will now use these ideas to exclude all generalized unitarity graphs other than the triangle graph. We begin by giving a more detailed analysis of how the integral (3.6) becomes a delta-function.

The analysis of the integral is complicated by the fact that different components of the momenta depend differently on Q. We can take care of this by associating with every momentum q a new momentum $\tilde{q}$, given by the linear transformation

$$\tilde{q}=Q^{-1}[e_1(e_1\cdot q)+e_2(e_2\cdot q)+Qe_+(e_-\cdot q)+Q^{-1}e_-(e_+\cdot q)]\,. \tag{3.8}$$

(This is the product of a dilatation and a Lorentz transformation.) The transformed external momenta are independent of Q:

$$\begin{aligned}\tilde{q}_1&=e_1+e_+\,,\\ \tilde{q}_2&=\tfrac{1}{2}(-e_1+\sqrt{3}e_2)\,,\\ \tilde{q}_3&=\tfrac{1}{2}(-e_1-\sqrt{3}e_2)-e_+\,.\end{aligned} \tag{3.9}$$

Likewise, with the internal momenta, the p's, we associate $\tilde{p}$'s, defined by the same transformation. It is easy to see that

$$d^4p = Q^4\, d^4\tilde{p}\,, \tag{3.10}$$

$$\delta(p^2) = Q^{-2}\delta(\tilde{p}^2) + \tilde{p}^{-2}\delta(Q^2)\,. \tag{3.11}$$

The origin of the delta-function is now clear. The relevant term in the integration must be the first term in eq. (3.11) for two of the delta-functions and the second term for one of them. That is to say, two of the $\tilde{p}$'s must be null and one non-null. It is easy to see which is which. At the q_1 vertex, the spacelike momentum $\tilde{q}_1$ becomes the sum of two momenta both directed forward in time. Thus one of $\tilde{p}$'s at this vertex must be non-null. Identical reasoning applies to the q_3 vertex. Thus the non-null momentum must be $\tilde{p}_2$.

All of this analysis extends to a general graph*. There are now more than three $\tilde{p}$'s, but since we only want one delta-function (because we only want to compute the residue of a single pole), only one of them can be non-null. Just as before, there must be at least one non-null $\tilde{p}$ emitted at the q_1 vertex and at least one non-null $\tilde{p}$ absorbed at the q_3 vertex. Thus there must be a single internal line (without any interactions along the line) joining these two vertices. This line carries a non-null $\tilde{p}$; all other $\tilde{p}$'s in the graph are null.

Now let us consider all the null $\tilde{p}$'s emitted at the q_1 vertex. We argue that these must all be parallel. For, if they were not parallel, they would represent massless particles moving in different directions. However, by classical kinematics, there would be no way for such particles to interact and change their directions of motion until they intersect one of the external currents. Thus, at most we could have two streams of particles, one heading towards the q_2 vertex and one heading towards the q_3 vertex. However, no null $\tilde{p}$ can go from the q_1 vertex to the q_3 vertex, for it would then be parallel to the unique non-null $\tilde{p}$. Identical reasoning applies to the null $\tilde{p}$'s absorbed at the q_3 vertex.

To summarize, we have found that the structure of the most general graph that contributes to the singularity we are studying is identical to the simple triangle graph, fig. 1, with one exception: the lines labeled p_1 and p_3 may represent not single massless particles, but streams of massless particles, moving together, and possibly interacting as they move along.

We can now use reasoning just like that which we used for fig. 3c to restrict the singularity associated with these streams. (Just as in the case of fig. 3c, the nice thing about this argument is that we can avoid looking in detail at the interactions of the particles as they move along.) If the singularity associated with the stream is stronger than a single pole, we have unacceptable long-range interactions between

* For the reader who is made nervous by the following blithe manipulation of singular functions, we give an alternative argument for the dominance of the triangle graph in appendix B, based upon a blithe manipulation of helicity zeros.

widely separated currents. If it is weaker than a single pole, we can discard the graph in our analysis. If it is exactly a single pole, we have a single-particle singularity.

Thus we have reached the desired result: the only graphs that contribute to the anomaly are triangle graphs, with massless particles moving around the triangle. Now we must look more closely at these particles; in particular, we must study their helicities.

4. Kinematic zeros and the importance of helicity one-half

Amplitudes for the emission or absorption of massless particles of non-zero helicity have kinematic zeros at vanishing particle energy. These are similar to the kinematic zeros that appear in ordinary potential scattering of spinless particles. Here the partial-wave amplitudes for non-zero angular momentum vanish at threshold, for the excellent reason that at threshold, where linear momentum vanishes, there can be no state of non-zero angular momentum. Likewise, in our case, for vanishing momentum there is no sensible definition of helicity, angular momentum along the direction of motion.

In S-matrix theory, there are definite rules for computing these helicity zeros. As we did for the Cutkosky rules, we shall simply state the rules here, and refer the reader interested in their derivation to the literature [6].

There are two sources of kinematic zeros:

Firstly, for every incoming or outgoing massless particle of energy E and helicity h there is a factor of $E^{|h|}$.

Secondly, with every incoming or outgoing massless particle there is associated a spinor transforming according to some representation of the homogeneous Lorentz group, O(3, 1). The helicity amplitude is multilinear in these spinors. To construct a Lorentz-invariant amplitude, it might be necessary to combine the spinors with the external momenta. This can lead to additional factors of E. If we label representations of O(3, 1) in the standard way, by two half-integers, the rules for associating spinors with particles are these: for every incoming particle of positive helicity, and for every outgoing particle of negative helicity, there is a spinor that transforms according to the representation $(|h|, 0)$. For the opposite signs of helicity, the spinor transforms according to the representation $(0, |h|)$.

For the special case of helicity $\pm\frac{1}{2}$, these rules should be familiar from field theories of neutrinos. The second rule is just the usual one for associating Weyl spinors with neutrinos or anti-neutrinos. The first rule is implicit in the normalization condition for these spinors, $U^{\dagger}U = 2E$. In the general case, the rules simply state that a particle of helicity h acts as if it were made of $2h$ non-interacting aligned neutrinos.

In the triangle graph, these rules are directly applicable to only one of the three vertices, the q_2 vertex, because only q_2 has all of its components small, proportional to Q. The other two external momenta, q_1 and q_3, have some components of order unity [see eq. (3.4)]. Nevertheless, they do have small squares, $q^2 = -Q^2$, so if we transform to the Breit frame, for example, all momenta will be of order Q and we can use our rules. However, when we transform back to our original frame, we may pick up a factor of Q^{-1} from the Lorentz transformation.

Thus, if the amplitude at the q_i vertex is proportional to E^{n_i} for small energies, we want

$$n_1 + n_2 + n_3 = 3 . \tag{4.1}$$

This will give us a factor of Q^3 for the three vertices, which, when combined with a factor of Q^{-1} for each of the two Lorentz transformations, will yield the desired factor of Q.

It will be convenient to orient our internal momentum along the arrows in fig. 1 and orient our helicities accordingly. Thus we consider the process at the q_1 vertex as the scattering of a particle of helicity h_2 off of a current into a particle of helicity h_3, rather than as the production of a pair with helicities $-h_2$ and h_3. Of course, either way of doing it gives the same power of Q, and by doing things this way we can do all three vertices just by cycling around the graph.

We can obtain an important constraint on the helicities by transforming to the Breit frame, in which the initial particle has momentum pointing in the positive z direction, and thus $J_z = h_2$, and the final particle has momentum pointing in the negative z direction, and thus $J_z = -h_3$. The current can only transfer $J_z = 0, \pm 1$. Hence

$$h_2 + h_3 = 0, \pm 1 . \tag{4.2}$$

The solutions to this equation fall into two classes. Either

$$h_2 h_3 \leqslant 0 , \tag{4.3a}$$

or

$$h_2 = h_3 = \pm\tfrac{1}{2} . \tag{4.3b}$$

These need to be analyzed separately.

For the first class, let us for definiteness assume that $h_2 \geqslant 0 \geqslant h_3$. We then have a spinor transforming according to $(h_2, 0)$ from the incoming particle and one transforming according to $(|h_3|, 0)$ from the outgoing one. We can combine these to make a spinor that transforms according to $(|h_2 + h_3|, 0)$; this is either $(0, 0)$ or $(1, 0)$. To this we must add one four-vector [transforming like $(\tfrac{1}{2}, \tfrac{1}{2})$] to construct a current [transforming in the same way]. Thus we obtain one power of Q from our second rule. If we add this to the powers of Q obtained from the first rule, we find,

$$n_1 = |h_2| + |h_3| + 1 , \tag{4.4a}$$

where we have written this equation in a form that is clearly also valid for $h_3 \geqslant 0 \geqslant h_2$.

For the second class, we have a spinor that transforms according to $(\frac{1}{2}, 0)$ for one particle and one that transforms according to $(0, \frac{1}{2})$ for the other. These combine to make $(\frac{1}{2}, \frac{1}{2})$; thus we need no extra powers of momentum, and

$$n_1 = |h_2| + |h_3| = 1 \, . \tag{4.4b}$$

The equations for n_2 and n_3 are obtained from these by cyclic permutation. We note that for all cases

$$n_i \geqslant 1 \, . \tag{4.5}$$

Thus the only way to satisfy eq. (4.1) is if

$$n_1 = n_2 = n_3 = 1 \, .$$

Thus, either

$$h_1 = h_2 = h_3 = 0 \, , \tag{4.6a}$$

or

$$h_1 = h_2 = h_3 = \pm\tfrac{1}{2} \, . \tag{4.6b}$$

The first possibility can be eliminated by *CPT* invariance. If we replace every helicity-zero particle by its *CPT* transform, the matrix elements of J^μ change sign. Thus the graphs with helicity-zero particles going around the triangle cancel in pairs. Only the second possibility remains, and the argument is complete.

5. Self-criticism

The flaws in our reasoning (or at least those we are aware of) fall into two classes.

First, we use the principles of analytic S-matrix theory, and the grounds on which these principles were established are far from solid, consisting in some cases of little more than perturbation theory and intelligent guessing. We do not think this is a serious problem; we believe the principles we have used are extremely plausible, even if not absolutely certain. However, it would indeed be very nice if they could be derived from the Wightman axioms, for example. We don't think this will happen soon, though; axiomatic field theory is notoriously difficult for theories with massless particles.

Secondly, and more seriously, in our analysis of singularities in sects. 3 and 4, we have had to continually assume that there are no singularities lurking in the vertices of our graphs. This assumption is not in general true. For example, if we consider the analogy with potential scattering made at the beginning of sect. 4, a partial-wave amplitude need not vanish at threshold if there is a threshold bound state; a dynamical singularity may cancel (or at least reduce the order of) a kinematic zero.

We think it might be possible to eliminate this possibility in the analysis of sect. 4. Our idea is best explained in a definite example. Suppose we have a vertex at which $h_2 = \frac{3}{2}$ and $h_3 = -\frac{1}{2}$. By eq. (4.4a) $n_1 = 3$; it was for this reason that we rejected this vertex in sect. 4. But actually, all we know from the analysis of kinematic zeros is that the matrix element of the current between the two massless particles is a kinematic factor that is proportional to Q^3 times an invariant function of Q^2. How do we know the invariant function does not have a pole at $Q^2 = 0$?

In this case, we can answer the question. If the invariant function had a pole at $Q^2 = 0$, it would imply that there would be a non-zero amplitude for the current to create a massless particle out of the vacuum. But by Goldstone's theorem, this would imply that the symmetry associated with the current would be spontaneously broken, and we have explicitly assumed that this is not the case.

We attempted to generalize this reasoning to take care of our problems, but we became lost in a dark wood of technical complications. This failure had led one of us (B.G.) to try a different approach, ignoring zeros and singularities altogether, and attempting simply to show that the putative wrong helicities are incapable of producing the proper tensor structure in the discontinuity of Γ. He is completing his investigation now and hopes to publish his results shortly.

We began this investigation at the 1980 summer session of the Aspen Center for Physics. We thank the Center and its staff for their support. We also thank Estia Eichten, John Preskill, and Lenny Susskind for conversations which strongly influenced the course of this work.

This research was supported in part by the National Science Foundation under grant no. PHY77-22864, and by the U.S. Department of Energy under Contract No. DE-ACO2-82ER40033.B000.

This paper is based on material submitted by one of us (B.G.) in partial fulfillment of the requirements for the Ph.D. degree at Harvard University.

Appendix A

NO NON-PERTURBATIVE CORRECTIONS TO THE ANOMALY

To show rigorously, independently of perturbation theory, that there is no correction to the anomaly equation in quantum chromodynamics would require first demonstrating, independently of perturbation theory, that quantum chromodynamics exists. This has not yet been done, so we must be less ambitious. We propose to show that there exists a cutoff procedure such that cutoff chromodynamics has no correction to the anomaly equation, independent of perturbation theory, for any value of the cutoff parameters. Thus, if one could show, as we can not, that the theory has a limit as the cutoffs are removed, one would know that this limit has no corrections to the anomaly.

We will work in the euclidean formulation of the theory, and in some fixed gauge (for example, axial gauge). We will place both an infrared and an ultraviolet cutoff on the gauge field. For the infrared cutoff, we will demand that the gauge field vanish outside some large box; for the ultraviolet cutoff, we will demand that the Fourier expansion of the gauge field within the box terminate at some large wave number. For our purposes, we will need neither an infrared nor an ultraviolet cutoff on the quark fields.

This procedure spoils both Lorentz invariance and translational invariance. However, it does not spoil chiral symmetry; J^{μ} is still conserved.

Now let us analyze the cutoff theory by first studying the dynamics of the quarks in a fixed gauge field. The gauge field is infinitely differentiable (because it is the sum of a finite number of Fourier components) and of compact support, so it is easy to show (for example, by the classic Fredholm methods used in abelian electrodynamics by Salam and Matthews [7]) that the quark propagator exists and that at short distances its leading singularity is the same as in free field theory.

Still working with a fixed gauge field, we construct Γ. This is exactly the triangle graph, with full propagators on the legs. Because the current is conserved, the anomaly is given by the leading short-distance singularity in Γ; this is given by the short-distance singularities in the propagators, and is thus the same as in free theory.

We now functionally integrate over the gauge fields. But since we have both an infrared and an ultraviolet cutoff, the functional integral is just an ordinary finite-dimensional integral. Thus if there is no correction to the anomaly in the integrand there is none in the integral.

This completes our hand-waving.

Appendix B

ALTERNATIVE TRIANGULAR ARGUMENT

In this appendix we give an alternative argument to the one given at the end of sect. 3, to show that only the triangle graph contributes to the singularity. This argument uses the rules for kinematic zeros given in sect. 4.

Let us consider a general graph that can contribute to the singularity, with all internal momenta null, forward-pointing, and collinear, as described in the paragraph following eq. (3.5). Let us refer to the three vertices at which the currents enter as "external vertices" and to the other vertices as "internal vertices". Let us consider a general internal vertex involving at least one particle of non-zero helicity. Since all momenta at such a vertex are null and collinear, we can make them all go to zero simultaneously by making an appropriate Lorentz boost; in this limit, the vertex must vanish. But the vertex is Lorentz-invariant; the only way it can do this is to vanish before the transformation. Thus each internal vertex must involve helicity-zero particles only.

By the arguments given in sect. 3, all internal vertices involving only three particles must vanish, or else we would obtain unacceptable long-distance interactions. (It is important to remember that a three-particle vertex is a number, a coupling constant, and thus there is no possibility of a cancellation when integrating over angles or such. If the unacceptable singularity vanishes, the vertex must vanish.) Thus each internal vertex must involve at least four particles.

At the external vertices, by the arguments given in sect. 4, there must be either particles of helicity zero or particles of helicity $\pm\frac{1}{2}$ (possibily with additional particles of helicity zero.) If every vertex has only helicity-zero particles, the contribution to the discontinuity vanishes, by the *CTP* argument given at the end of sect. 4. Thus at least one vertex must involve a pair of particles of helicity $\pm\frac{1}{2}$. Since these particles are not allowed to interact at internal vertices, they must appear at all the external vertices to keep the graph connected. Also to keep the graph connected, there must be at least two external vertices at which helicity-zero particles appear.

To summarize, we have found that a general graph contributing to the singularity consists of the fundamental triangle of helicity $\pm\frac{1}{2}$ particles, decorated by helicity-zero particles. These helicity-zero particles must be created at some external vertex and destroyed at some other external vertex. Between their creation and destruction, they are allowed to interact with each other, but only through vertices involving at least four particles. They are not allowed to interact with the particles of helicity $\pm\frac{1}{2}$.

We have computed in the text that the discontinuity is a homogeneous function of Q of order minus one, if there are no internal helicity-zero particles. We are now in a position to compute the order if there are.

Let us denote the order of the discontinuity by N, the total number of vertices in the graph (internal and external) by V, the number of internal lines (of any helicity) by I, and the number of independent closed loops by L. Then, by elementary power counting,

$$N = 4L - 2I + 1\,,$$

where the last term comes from the kinematic zeros of the external vertices, as computed in sect. 4. By standard reasoning,

$$L = I - V + 1\,.$$

Thus,

$$N = 2I - 4V + 5\,.$$

Every internal line has two ends which terminate on vertices. Every internal vertex has at least four internal lines terminating on it. The three external vertices have at least eight internal lines terminating on them, six lines for particles of helicity $\pm\frac{1}{2}$ and at least two lines for particles of helicity zero. Thus,

$$2I \geqslant 8 + 4(V-3)\,.$$

Hence

$$N \geqslant 1\,.$$

But this means the graph contributes nothing at all to the discontinuity, since there is no homogeneous function of positive order with support restricted to $Q^2 = 0$.

References

[1] G. 't Hooft, *in* Recent developments in gauge theories, ed. G. 't Hooft et al. (Plenum, N.Y., 1980)

[2] R. Eden, P. Landshoff, D. Olive and J. Polkinghorne, The analytic S-matrix (Cambridge University Press, Cambridge, 1966)

[3] Y. Frishman, A. Schwimmer, T. Banks and S. Yankielowicz, Nucl. Phys. B177 (1981) 157;
A.D. Dolgov and V.I. Zakharov, Nucl. Phys. B27 (1971) 525

[4] J. Bell and R. Jackiw, Nuovo Cim. 60A (1969) 47;
S.L. Adler, Phys. Rev. 177 (1969) 2426

[5] S. Weinberg and E. Witten, Phys. Lett. 96B (1980) 59

[6] G. Cohen-Tannoudji, A. Morel and H. Navelet, Ann. of Phys. 46 (1968) 239

[7] A. Salam and P. Matthews, Phys. Rev. 90 (1953) 690